Statistica Neerlandica (2012) Vol. 66, nr. 3, pp. 217–236
doi:10.1111/j.1467-9574.2012.00530.x

'All models are wrong...': an introduction to model uncertainty

Ernst Wit*

Johann Bernoulli Institute, University of Groningen, Nijenborgh 9, 9747 AG Groningen, The Netherlands

Edwin van den Heuvel†

Epidemiology, University of Groningen UMCG, Hanzeplein 1, 9713 GZ Groningen, The Netherlands

Jan-Willem Romeijn‡

Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands

In this article, we introduce the concept of model uncertainty. We review the frequentist and Bayesian ideas underlying model selection, which serve as an introduction to the rest of this special issue on 'All models are wrong...', a workshop under the same name was held in March 2011 in Groningen to critically examined the field of statistical model selection methods over the past 40 years. We briefly introduce the philosophical debate that is concerned with model selection. We present the results of a questionnaire that was distributed under the participants of the workshop, showing that the field has not yet reached a comforting consensus and is still in full swing.

Keywords and Phrases: model selection, model uncertainty, Bayesian, frequentist, foundations, confirmation theory, questionnaire.

1 Introduction

On Tuesday morning March 16, 1971, Hirotake Akaike sat in the metro in a Tokyo suburb with an impending deadline for a paper. On that journey, the general idea behind the future Akaike's information criterion (AIC) suddenly dawned on him. Four decades later, the model selection landscape has become much more complex, involving themes such as sparsity, high-dimensionality, explanatory versus predictive ability, among other issues. To reflect on the achievements over this period, the authors organized exactly forty years later a workshop on 'All models are wrong: model uncertainty and selection in complex models' held from 14–16 March, 2011 in

*e.c.wit@rug.nl
†e.r.van.den.heuvel@umcg.nl
‡j.w.romeijn@rug.nl

Groningen. This special issue of *Statistica Neerlandica* brings together statisticians, philosophers and quantitative scientists to address the complexity of model uncertainty, cutting-edge strategies to deal with model uncertainty and common (or not so common) pitfalls in model selection. All of the authors were contributors to the workshop and they were asked to also pay attention to foundational issues.

A statistical model is typically defined as a collection of probability measures on some outcome space. It comes very naturally to a statistician that all, except perhaps one, of these measures are 'wrong'. Traditionally, selection of one of these models was based on fit, i.e. the distance from the data to the model. 'Fit' can be defined in many ways, such as via some loss function, such as least squares, or maximum likelihood. Initially, it were semantic requirements of interpretation, that drove early statisticians such as Karl Pearson and Ronald A. Fisher to discard 'best fitting' models for simpler, but less well-fitting models. Hypothesis testing was essentially introduced as a tool to trade-off fit for simplicity.

Within regression modelling, probability measures are indexed by regression parameters. Regression semantics often involves answering questions such as whether a variable should be included in the model. This, following Fisher's lead, became the subject of hypothesis testing in the fifties and sixties. There was also criticism. Freedman's paradox (Freedman, 1983) showed how repeated selection of variables by testing distorts nominal significance levels and gives no guarantee against including unrelated variables. In fact, the rather unspecific question 'what is the best model?' turned out to have more than one meaning: (i) which modelling procedure will, with sufficient data, identify the true model?, or (ii) based on the data, which model lies closest to the true model? Despite the apparent similarities between these questions, they turned out to be quite different and surprisingly incompatible.

2 AIC and BIC: between proximity and truth

In this section, we will deal with the methodological and technical aspects of model uncertainty. First, we focus on frequentist methodologies, which are based on finding models close to the truth and involve either implicitly or explicitly some consideration of model complexity. Then we consider orthodox Bayesian methods aimed at obtaining the posterior distribution on the model space. This means that they implicitly presuppose that the integral of all model posterior probabilities equals one, i.e. that the Bayesian believes that one of the models under consideration is the true model.

2.1 Akaike's information criterion

Akaike's information criterion did not completely fall out of the sky. In the 1960s, many applied statistician grabbled with variable selection as one of the primary problems within statistical modelling. In 1964 and 1966, C.L. Mallows gave two talks

at IMS and ASA meetings on a visual aid for variable selection. Mallows (1973) presented the method formally as a way to minimize the expected prediction error, which was also the point of departure of Akaike. Interestingly, Mallows (1973) warns against using 'his' C_p method as a model selection tool, instead of a method for selecting a set of models with good predictive performance.

Akaike's contribution was to look beyond the regression model and to make the method available in a general likelihood setting. In the usual setting of n independent observations, we can define a *statistical model* $\mathcal{M}_i = \{\mathcal{P}_\theta \mid \theta \in \Theta_i\}$ as a collection of measures on the data $Y \in S^n$ for some outcome space S, indexed by a p_i-dimensional parameter space Θ_i. When we consider several competing models, we are, in fact, considering competing *collections* of measures on the data. Let's consider a finite number k of such collections, i.e., $\mathbb{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$, each indexed by their own finite dimensional parameter space Θ.

Akaike proposed to minimize a quantity, called the AIC, across $\mathbb{M}$, where the AIC is defined as

$$\mathrm{AIC}(\mathcal{M}_i) = -2 \log P_{\hat{\theta}_i}(Y) + 2p_i, \tag{1}$$

where $p_i = \dim(\Theta_i)$. The idea of the AIC is to select the best approximating model to the unknown true data generating process, in the sense of the Kullback–Leibler (KL) divergence. As we do not know the measure that corresponds to the true data generating process, we have to estimate this divergence. In the derivation of the AIC, i.e. the estimation of the KL divergence, one uses several asymptotic identities. Especially for small samples, these approximations can be too liberal, in the sense of favouring larger models. Instead, small sample results have been derived in several special cases. For example, for the normal linear regression model the so-called *corrected* AIC, or AIC_c, is given as

$$\mathrm{AIC}_c(\mathcal{M}_i) = -2 \log P_{\hat{\theta}_i}(Y) + \frac{2p_i}{n/(n - p_i - 1)},$$

where the term $n/(n - p_i - 1)$ is a first order bias correction in the estimation of the KL divergence, that goes to 1 when the number of observations goes to infinity. This correction seems to 'work' quite well also in other circumstances.

2.1.1 Derivation of the AIC

Suppose the data Y is generated from the true density g. Let $\mathcal{M}_i$ be our parametric model with parameter space Θ_i, a compact subset of $\mathbb{R}^{p_i}$. Let $\hat{\theta}_i$ be the maximum likelihood estimator (MLE), then $\hat{\theta}$ minimizes the Kullback–Leibler (KL) divergence between the fitted and the true model

$$KL(\theta) = \int g(y) \log g(y)\, \mathrm{d}y - \int g(y) \log f(y; \theta)\, \mathrm{d}y,$$

where $f(\cdot; \theta)$ is the density associated with $\mathcal{P}_\theta$. If we were lucky to select $f(\cdot; \theta) = g$, then the divergence would be zero. Simply selecting the model $\mathcal{M}_i$ that minimizes the

KL divergence, however, is not an option as bigger models have a non-decreasing divergence compared to nested alternatives. Minimizing $E_g(KL(\hat{\theta}_i))$ across i does not suffer from this problem, as, roughly, the MLE $\hat{\theta}$ and the average KL divergence are calculated on different datasets. The problem is that this quantity is unobservable and needs to be estimated. First of all $C = \int g(y) \log g(y)\, \mathrm{d}y$ is unobservable, but this can be ignored, because it is constant across all Θ_i. This leaves us with estimating,

$$l_i = E_g \int g(y) \log f(y; \hat{\theta}_i)\, \mathrm{d}y,$$

where the expectation is over the data used for estimating $\hat{\theta}_i$. The natural estimator

$$\hat{l}_i = \log f(y_o; \hat{\theta}_i(y_o)),$$

where y_o is the observed data, typically overshoots its target l_i by a quantity given as

$$E(\hat{l}_i - l_i) \approx p_i^*,$$

where

$$p_i^* = \mathrm{Trace}\,(J^{-1}K),$$

where both J and K are forms of the expected observed Fisher information using model $\mathcal{M}_i$

$$J_i = E_g\left(\frac{\partial^2 \log f(Y, \hat{\theta}_i)}{\partial\theta\partial\theta^t}\right), \quad K_i = V_g\left(\frac{\partial \log(f(Y, \hat{\theta}_i)}{\partial\theta}\right).$$

These $p_i \times p_i$ matrices are identical if $g \equiv f(\cdot, \hat{\theta}_i)$, in which case $p_i^* = p_i$. One could try to estimate the matrices J_i and K_i, as is done for example in the Takeuchi Information Criterion (TIC) by replacing J_i and K_i by their empirical counterparts. However, this can be very unstable. Typically, it is better to use p_i directly, which leads to the following estimator of the KL divergence,

$$\widehat{KL}_i = C - \hat{l}_i + p_i,$$

where $C = \int g(y) \log g(y)\, \mathrm{d}y$ is an irrelevant constant that does not depend on i. Taking out C and multiplying the above equation by 2 gives us the AIC as defined in Equation (1).

2.1.2 *Van der Linde's model complexity*

Interestingly, in the derivation of the AIC, a notion of model complexity appears as a relevant term in determining the best model. Apparently, if one 'penalizes' the likelihood by some model complexity term, one obtains a measure of how close a statistical model approaches the true data generating measure. This opened up a important line of research in ideas about model complexity, or model *degrees of freedom*. Important developments in general *covariance penalty theory* (Friedman, Hastie and Tibshirani, 2001; Efron, 2004) showed that the degrees of freedom can

be generalized as

$$gdf = \sum_{i=1}^{n} \mathrm{cov}(Y_i, \hat{\mu}_i(Y)),$$

where μ is the natural parameter associated with the loss function. For the squared loss function, this is the mean parameter and in the case of ordinary multiple linear regression modelling, this results in $gdf = \mathrm{Tr}(X(X^tX)^{-1}X^t) = p$, where p is the rank of the matrix X.

Model complexity is, however, itself a complex issue, for which various points of view exist. Angelika van der Linde explains in 'A Bayesian view of model complexity' in this volume that there are a variety of definitions for model complexity. Roughly speaking, she identifies two forms of model selection targets: *average* and *representative* targets. Average targets correspond to using mutual information as measure of model complexity, which occurs naturally in Bayesian modelling. Representative targets of model complexity correspond to using the KL-divergence with a representative density, as is done in the AIC. In her article, she gives an extensive review of various measures of model complexity. In the end, she favours the information theoretic ideas of model complexity, especially in an new era of statistics, where neural networks, mixtures, reduced rank regression, hidden Markov models etc. do not satisfy the commonly assumed regularity conditions (like positive definiteness of the Fisher information matrix) underlying the AIC and BIC.

2.1.3 Claeskens' FIC

As several authors have pointed out already, a model selection algorithm cannot be optimal both for prediction accuracy and for variable selection consistency. A universal 'best' model selection criterion is therefore impossible. Claeskens argues in her article that this implies that one has to choose a model selection *focus*. She defines a focussed model selection criterion (FIC) that aims to minimize the mean squared error of a particular parameter estimate, such as the mean, the variance, or particular covariate parameters.

Despite formulating the information criterion as a minimization problem of the mean squared error, there is some connection with the AIC. It can be shown that if the focus is a single parameter, then the FIC and the AIC are identical (Claeskens and Hjort, 2008). This is a rare case, but it does show that the FIC, just like the AIC, is more interested in prediction accuracy, rather than variable selection consistency. In fact, in her contribution, 'Focused estimation and model averaging with penalization methods, an overview', Claeskens applies the FIC to a *non-sparse* approximation to the lasso. The lasso (Tibshirani, 1996) is a penalized regression method that automatically select variable depending on the penalty parameter. Claeskens argues that 'the decision to set variables to zero (hence to not select them for inclusion in the model) should come from the value of the FIC (which combines the loss function of choice with the specific focus to estimate)' and not 'be determined by the penalty itself, without taking the

focus and loss into account.' This is a philosophy of science that is inspired by the use of models in the social sciences, psychology and economics. Empirical models are tools to achieve some end, which, in the light of Popper, Kuhn and Lakatos, is rarely 'truth itself'. Variable selection consistency is a concept entirely foreign to such pragmatic model builders. They are happy to ignore variables if it is useful, but for no moment do they actually believe that the variables have no influence at all. As statistician Gelman (2011) argues:

> in the sort of social science problems I study, there are no true zeroes except by design or through a natural experiment, and I do not see the point of statistical methods that attempt to discover from data conditional independence patterns that cannot exist... I follow Popper in believing that a model can be rejected, never accepted. I will go even further and say that, realistically all my models are wrong.

2.2 *Bayesian model uncertainty*

Contrary to the frequentist position, the orthodox Bayesian framework is interested in nothing less than the truth itself. Bayesians are not naive and famously allow for uncertainty in their opinion about the truth. This uncertainty can stretch very far: a Bayesian may not only be uncertain about the parameter values, but even about the models themselves. In fact, this is how Bayesians view model uncertainty: as determining the posterior distribution $p(y \mid \mathcal{M})$ on the model space $\mathbb{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_k\}$. Hastie and Green argue that the quantities $p(\mathcal{M} \mid y)$ are the appropriate probabilities particularly for subsequent inferential activities, such as prediction

$$p(y^+ \mid y) = \sum_{\mathcal{M} \in \mathbb{M}} p(y^+ \mid \mathcal{M})\, p(\mathcal{M} \mid y),$$

where y^+ is some new data point, or estimation

$$E(F \mid y) = \sum_{\mathcal{M} \in \mathbb{M}} \int F(\theta)\, p(\theta \mid \mathcal{M})\, \partial\theta p(\mathcal{M} \mid y),$$

for some function F of the parameters with 'constant' interpretation across the various models in $\mathbb{M}$.

2.2.1 *Friel and Wyse's evidence*

Given a class of models $\mathbb{M}$, each one with its own parameter space $\Theta_{\mathcal{M}}$, where $\mathcal{M} \in \mathbb{M}$, the first task of a Bayesian is to ascertain the prior distribution on $\mathbb{M} \cup \bigcup_{\mathcal{M} \in \mathbb{M}} \Theta_{\mathcal{M}}$. After seeing the data, the Bayesian will update her prior and in principle would be able to hand over the marginal posterior distribution on $\mathbb{M}$. As a distribution, it would reflect her true model uncertainty. She could also use this posterior to perform some 'unbayesian' activities: she could consider the model $\mathcal{M}$ that maximizes her posterior, or compare two model via a so-called Bayes' factor, $p(\mathcal{M}_1 \mid y)/p(\mathcal{M}_2 \mid y)$.

The posterior model density involves the usual prior and likelihood contributions and can be written as

$$p(\mathcal{M} \mid y) = \frac{p(\mathcal{M})\, p(y \mid \mathcal{M})}{p(y)} \tag{2}$$

$$= \frac{p(\mathcal{M})}{p(y)} \int_{\Theta_{\mathcal{M}}} p(y \mid \mathcal{M}, \theta)\, p(\theta \mid \mathcal{M})\, \partial\theta \tag{3}$$

The quantity $p(y \mid \mathcal{M}) = \int_{\Theta_{\mathcal{M}}} p(y \mid \mathcal{M}, \theta)\, p(\theta \mid \mathcal{M})\, \partial\theta$ is what is described in the Friel and Wyse contribution as the *marginal likelihood*, *integrated likelihood* or simply as the *evidence*. Especially the last name is suggestive, since it is the only element in Equation (2) that allow the data to favour one or another model, since $p(y)$ is constant and $p(\mathcal{M})$ is the prior, chosen irrespective of the data.

Friel and Wyse dedicate their article in explaining the various ways that the evidence can be calculated in practice. This is a non-trivial task as the integration in Equation (3) for moderately complex models can be across a large sample space $\Theta_{\mathcal{M}}$.

2.2.2 *Bayesian information criterion*

The simplest method for calculating the integral is based on the Laplace approximation and was proposed in Schwarz (1978). The Laplace approximation states that for a function f with a global maximum at x_0, we have for large C

$$\int_{-\infty}^{\infty} e^{Cf(x)}\, \mathrm{d}x \approx \sqrt{2\pi} C\, |f''(x_0)|\, e^{Cf(x_0)}.$$

If we take a flat prior on $\Theta_{\mathcal{M}}$, the evidence can be approximated for $\mathcal{M} \in \mathbb{M}$ as

$$\begin{aligned} p(y \mid \mathcal{M}) &= \int_{\Theta_{\mathcal{M}}} e^{n\bar{l}(\theta)}\, \partial\theta \\ &\approx e^{l(\hat{\theta})} \int_{\Theta_{\mathcal{M}}} e^{-\frac{1}{2}(\theta-\hat{\theta})^t n \frac{\partial^2}{\partial\theta^2} \bar{l}(\hat{\theta})(\theta-\hat{\theta})}\, \partial\theta \\ &\approx e^{l(\hat{\theta})} (2\pi)^{-p/2} n^{-p/2} \left| \frac{\partial^2}{\partial\theta^2} \bar{l}(\hat{\theta}) \right|^{-1/2}, \end{aligned}$$

where $\bar{l}(\theta) = l(\theta)/n$ is the mean log-likelihood for model $\mathcal{M}$ and p the dimensionality of the parameter space $\Theta_{\mathcal{M}}$. Schwarz (1978) proposed to ignore, for large n, the terms that do not depend on n, resulting in

$$p(y \mid \mathcal{M}) \approx e^{l(\hat{\theta})} n^{-p/2}, \quad \text{for large } n.$$

If, furthermore, one is willing to put a flat prior on $\mathbb{M}$, then the posterior model probability for large n can be written as

$$p(\mathcal{M} \mid y) \approx c e^{l(\hat{\theta})} n^{-p/2} / p(y), \quad \text{for large } n.$$

By applying an arbitrarily monotone decreasing transformation to the above and ignoring the constant terms c and $p(y)$, we get

$$\mathrm{BIC}(\mathcal{M}) = -2l(\hat{\theta}) + p\log(n),$$

which is named the *Bayesian information criterion* (BIC). Minimizing the BIC corresponds to maximizing the posterior model probability for a large number of observations and can be thought of as a way of selecting a model. For a Bayesian, however, it would be more appropriate to consider the model weights $W(\mathcal{M})$,

$$W(\mathcal{M}) = e^{-\mathrm{BIC}(\mathcal{M})/2},$$

which in their rescaled form exactly correspond to the posterior model probabilities themselves,

$$p(\mathcal{M}_0 \mid y) = \frac{W(\mathcal{M}_0)}{\sum_{\mathcal{M}\in\mathbb{M}} W(\mathcal{M})}.$$

2.2.3 Hastie and Green's reversible jump MCMC

In their article, Hastie and Green aim to obtain the posterior model probability $p(\mathcal{M} \mid y)$ in another way. They use the idea that it is possible to sample from any distribution p, if this distribution is the stationary distribution associated with some Markov Chain. Green (1995) has shown that this is possible in a fundamental way even for model probabilities by ingeniously constructing a transdimensional Markov Chain. By Monte Carlo sampling from this Markov chain, the relative frequencies of visiting each model $\mathcal{M} \in \mathbb{M}$ converges to the posterior model probability $p(\mathcal{M} \mid y)$. Despite this attractively simple sounding summary, the devil is definitely in the details. Most of the Hastie and Green contribution is about implementational issues regarding the Markov chain. The usual Monte Carlo Markov Chain (MCMC) issues, such as sensible priors, effective move proposals, sufficient mixing and convergence, are at least twice as difficult in a so-called reversible jump MCMC, used for walking around the model space.

3 Model selection as a philosophical topic

As introduced in the foregoing, statistical model selection is first and foremost a topic within mathematical statistics. But foundational questions are never far off. If all models are wrong, in the sense that models never tell the whole truth but at best represent fragments of it, then what motivates our focus on particular fragments? How does this relate to the multitude of measures for comparing models? And why rely on simple models rather than complex ones? The past twenty years or so have seen a lively philosophical debate over such questions, and to most statisticians these questions will sound familiar too. In what follows we provide a quick introduction

to the foundational side of model selection, and we briefly comment on foundational aspects of the papers in this special issue.

3.1 Philosophical perspectives on model selection

When characerised in sufficient generality, it becomes apparent that statistical model selection answers a traditional philosophical question. Model selection concerns the choice of a model, a theoretical framework that accommodates and predicts empirical data. A model selection tool fixes the criteria for choosing among models, and thereby specifies what aspects of accommodation and prediction are valued, and to what degree. Model selection theory can therefore be considered part of *confirmation theory*, a discipline within the philosophy of science that is concerned with norms for the relation between theory and data.

Much of traditional confirmation theory has focussed on an even more basic statistical topic, namely the evaluation of hypotheses against the background of a statistical model (Romeijn and van de Schoot, 2008; Romeijn, 2011). Inductive logic (Woods and Hartmann, 2011) formal learning theory (Kelly, 1996) so-called Bayesian confirmation theory (Earman, 1992), and other perspectives involved in the debate over paradoxes of confirmation (e.g. Eells and Fitelson, 2000; Fitelson, 2006) provide accounts of the support that data lend to a hypothesis within a given probabilistic setting. The topic of model selection is markedly different. It is not so much aimed at the evaluation of a particular hypothesis in a given setting, but rather at the evaluation of the setting itself. Whereas, for example, Carnap's inductive logic (Carnap, 1952) evaluates predictions on the assumption that particular patterns in the data are salient, model selection is aimed at evaluating the focus on particular patterns in comparison to other patterns in the data.

In short, model selection concerns not so much statistical procedures and inferences, but rather their assumptions and starting points. Consequently, foundational questions about model selection cannot be answered solely by paying close attention to mathematical details in the procedures and inferences involved. They require us to confront fundamental issues in statistics, and to look at the wider context in which model selection is used. Perhaps the most central issue here is the so-called reference class problem (Hájek, 2007). The probability of an event is associated with a population of which the event is a subset. The reference class problem states that many such populations typically do exist. If Anna is a smoking vegetarian who studied maths and does Yoga, then what population should we choose to determine the probability that she will eventually develop Alzheimer's disease? Or, to put the question in model selection form, what variables should we include in the statistical model that will be used to predict Anna's old age? In the face of this context sensitivity, our intuition that there is such a probability seems very hard to uphold.

One way of looking at model selection is that it is aimed at resolving the problem of the reference class. There are again different ways of looking at the solution that model selection provides. One prominent way emphasizes the pragmatic aspects,

taking model selection as a trade-off between the goals of accommodating old data and predicting new (Forster and Sober, 1994; Hitchcock and Sober, 2004) or, equivalently, of balancing bias and variance, or, again equivalently, of weighing complexity against fit. On the one hand, we want to avoid including too many variables in the model, because the estimation of the parameters in the model will in that case become very variable. In other words, we want to avoid accommodating the historical data too much and thereby getting unreliable predictions on future data. On the other hand, we also want to avoid including too few variables in the model, because in that case prediction will be heavily biased. That is, we want to avoid accommodating the data too little and thereby making our predictions inaccurate. But there are other ways of looking at model selection. We might for instance maintain that model selection tools are an instrument for getting at the true distribution, and that their pragmatic value is a mere consequence of their ability to hone in on the truth.

However one looks at model selection, one theme is always prominent and continues to capture the interest of philosophers of science: simplicity. Especially appealing is the idea that model selection tools provide an independently motivated formalization of simplicity (Forster, 2000; Kelly, 2007), a concept that has been present in philosophical discussions ever since Ockham's *Numquam ponenda est pluralitas sine necessitate*[1] but that seems to have eschewed analytic treatment until only a few decades ago. There are good reasons to evaluate this idea critically. First, while simplicity as dimensionality indeed follows from independently motivated epistemic goals, like proximity to the truth (AIC) or probability of truth (BIC), other approaches such as the Minimum Description Length principle (Rissanen, 1983; Grünwald, Myung and Pitt, 2005) assume a concept of simplicity at the outset. Second, there are several aspects of models that are intuitively associated with their simplicity but that do not register in traditional model selection tools (Romeijn, 2012). Third, and most obviously, there are many examples of the role of simplicity in science that cannot be fitted into the mold of statistical model selection.

Nevertheless, it does seem that statistical model selection presents ample opportunity for exploring and structuring the concept of simplicity (Dowe, Gardner and Oppy, 2007). And more generally, it seems that the potentials of model selection as an integral part of confirmation theory are far from exhausted (Henderson *et al.*, 2010). Again closer to the foundations of statistics there are many aspects of model selection presently left unexplored (Chakrabarti and Ghosh, 2011). For all these reasons, we expect the foundations of model selection to be a lively field of inquiry for decades to come.

3.2 Philosophical perspectives in this special issue

The papers of Longford and van der Linde both make general points on the theory of model selection. Claeskens considers particular model selection tools, while Green and Friel are both concerned with the computation and computational tractability

of model selection problems. Finally, Wenmackers confronts statistical model selection as it appears in the social sciences with the choice of models in experimental physical science. The first two and the last of these will be given special attention because they are most naturally related to the themes raised in the foregoing.

3.2.1 Theoretical views

Longford's paper brings to the fore an aspect to model selection that has received only cursory attention up till now. Model *selection* tools are used to decide between models. Longford's argues that the decision to use one model to the exclusion of others need not be the result of model selection. Instead, Longford proposes to use the so-called composition of estimators, essentially replacing the estimator of the best model by a weighted linear combination of the estimators from all models, using the quality of the estimators, typically expressed by the mean squared error, as weights. One might say that this proposal is the frequentist equivalent of Bayesian model averaging. The approach of Longford further highlights that model selection is intimately connected to decision problems and that it need not lead to the decision to use a single model. In this sense his paper is also related to foundational work in statistics that emphasizes the necessity of a decision–theoretic component in any theory of statistics (Savage, 1951; Seidenfeld, 1979).

Van der Linde's paper addresses a question that is directly related to the foregoing discussion on the trade-off between variance and bias. Van der Linde observes that within Bayesian approaches to model selection, we find many different expressions for the complexity, or the *effective number of parameters*, of a model. This observation motivates the introduction of an abstract notion of complexity that may unify these different expressions. The abstract notion, to which van der Linde appeals, stems from information theory: the mutual information of data and model parameters. The joint density over data and parameters that serves as input to the mutual information comes in two versions, namely prior and posterior, depending on whether the data are those already obtained and the distribution over parameters is the prior, or else the data are those over which the model expresses expectations and the distribution over parameters is the posterior. Moreover, by replacing the Bayesian model averages by the best-fitting distributions from the models, frequentist approaches can be considered in the same information–theoretic perspective. An especially appealing aspect of this perspective on model complexity is that it emerges from taking the marginal entropy of the distribution over data as an expression of model performance. The marginal entropy naturally decomposes into a goodness-of-fit term and a mutual information term, suggesting that the information–theoretic perspective motivates a specific balance between bias and variance.

This latter point calls for some further philosophical reflection. Van der Linde does not claim there is a one-size-fits-all answer to the question of how to trade bias against variance. She shows herself sensibly sensitive to the fact that no such answers exist and that statistical procedures are essentially context-dependent. Nonetheless,

information theory seems to provide a framework for unifying different expressions of model complexity, both for frequentists and Bayesians and both for prior and posterior versions of model performance. We may expect that the question of how to balance bias and variance is properly structured by this framework, so that everyone can at least agree on what considerations determine the balance. A question that is open for exploration is how this relates to philosophical views on this trade-off. In particular, it may be argued that the existence of this common framework points to the trade-off being not purely pragmatic, but, at least partially, a matter of principle.

3.2.2 Applications and computation

In contrast to the unifying perspective of van der Linde, the paper by Claeskens illustrates the wide variety of applications of model selection, thereby suggesting its ultimately pragmatic character. Claeskens' paper deals with model selection under a further specification of the aims of the model, namely of getting a particular estimate right. As Claeskens shows, the specification of such a focus for the model selection procedure may be favourably employed when dealing with, in this case, high-dimensional data.

The papers by Hastie and Green and Friel and Wyse are both concerned with computational aspects of Bayesian model selection. Despite present appearances to the contrary, philosophical perspectives are certainly not insensitive to computational worries, as witnessed by the interest in network representations and local computation (Haenni *et al.*, 2010) and the growing literature on the epistemic status of simulations (Morgan, 2005; Frigg and Reiss, 2009). For the purpose of this introduction, however, including these philosophical perspectives will lead us too far afield.

3.2.3 Broader reflection

Wenmackers and Vanpoucke discuss the phenomenon of a scientific model in the context of material science, involving both empirical and strictly theoretical structures in their account. They analyze the evaluation of these models in experimental physics. Their discussion provides some valuable observations and reminders for both philosophers and practitioners of model selection. One of their observations is simply that scientific models are always abstract and idealized versions of the target system, thereby defusing the claim that all models are *wrong*. The key is not to be *too* wrong. A second observation is that considerations that lead to the preferred level of abstraction are mostly pragmatic in nature, once again highlighting the context dependence of balancing bias and variance. A third observation pertains to scientific realism and stands in stark contrast to the second one. Moreover, it connects in interesting ways to the topics discussed in the first subsection. Following the theoretical viewpoint of most material scientists, we might portray statistical model selection as getting at the true chances of events. Accordingly, we might motivate

the trade-off between bias and variance by its truth-conduciveness rather than by their pragmatic value. Of course, such a perspective hinges on there being a sensible way of spelling out the notion of true chance.

Perhaps most prominently, the paper by Wenmackers and Vanpoucke reminds us of the fact that the internal variance of the statistical model does not capture all of our uncertainty on how the model relates to the world out there. Following the detailed description of the experimental procedures, it becomes apparent that at least some of our uncertainty remains hidden from view, because it is caused by uncertainty over assumptions that were made to obtain clean data sets. The very same thing can of course be said about statistics. It is a challenging suggestion, especially in a statistics journal, that quantitative error bars may well lead to a false sense of security. However comprehensive the model, some uncertainties will not be captured in probabilistic terms and hence must be expressed qualitatively.

4 Questionnaire

During the organization of the workshop, the idea of a questionnaire arose to find out whether professionals interested in model selection share a general consensus on this topic or whether they have substantial different views. Traditionally, model selection was performed on the basis of hypothesis testing in combination with forward, backward or stepwise approaches. With the introduction of Akaike's information criteria, hypothesis testing as a model selection tool has been shown to be problematic, but it has not been eliminated from the traditional toolbox of the scientist. In the field of model selection many different information criteria have been developed to improve upon the work of Akaike, and also Bayesian approaches like model averaging was introduced in section 2. Thus besides a traditional controversy in model selection on hypothesis testing another form of controversy between frequentists and Bayesians has entered this domain of research.

Eleven multiple choice questions were constructed to obtain the opinions and views of the community on model selection. All participants, including speakers, registered for the workshop were sent the questionnaire by email and were asked to explain their answers as much as possible to each question. Three questions (1, 4, and 9) asked about personal favorites, three questions (2, 6, and 10) were more of a general nature, two questions (5 and 7) asked about the purpose of some techniques, and finally three questions (3, 8, and 11) were more of a technical nature, see Appendix A. Question twelve gave the participants the opportunity to formulate their own questions and answers. In total 110 participants received the questionnaire, but only 16 persons returned their answers to us and even within this group not all questions were answered by everyone or explained in much detail. Some characteristics of the responders are given in Table 1.

Although it is a small group of respondents, the diversity in characteristics is relative large. Indeed, they include invited and contributed speakers, poster presenters,

Table 1. Characteristics of the participants that filled in the questionnaire

Type of participant (n)	Age (n)	Gender (n)	Answering questionnaire (n)	Workshop changed my mind (n)
Invited speaker (3)	≤ 30 (3)	Male (8)	Before workshop (7)	Yes (6)
Contributed speaker (5)	(30, 60] (9)	Female (6)	After workshop (8)	No (4)
Poster presentation (2)	>60 (3)	More subtle (0)	NA (1)	NA (6)
Attendee (5)	NA (1)	NA (2)		

and attendees. Furthermore, all ages were represented as well as both sexes. About 50% of the respondents answered the questionnaire before the workshop and one fourth changed their mind on model selection concepts endorsed by the workshop. The answers to the questions of these respondents are listed in Appendix A by a frequency behind the answers.

4.1 Personal favorites

Based on the personal choices of the respondents to questions 1, 4 and 9, the group of respondents seems quite heterogeneous. Indeed, the respondents see themselves as different types of persons and they prefer a wide variety of measures for model comparison. Some respondents strongly discard model averaging on the basis of having no justification of model weights and they believe that model averaging may not be applicable in each context. One remark was that model averaging is prohibited by the publish-or-perish culture in science. Most respondents felt that the choice between model averaging and selection depends on the research question and they did not classify themselves to just one of the two groups. Some respondents mentioned that a mechanical approach towards selection and averaging should be avoided, because the 'best' model may not be the right choice in some circumstances. The model averaging persons selected their option since they believe in posterior model probabilities or are true Bayesians. Some selected alternative measures for model comparison; the mean squared error and false negatives. The Kullback–Leibner distance was defended by the fact that it relates to likelihood, which may be considered the key measure for parametric inference. Some respondents made a particular choice for themselves, but argued that they felt comfortable with some of the other measures as well, like the Kullback–Leibner distance. An important remark towards the choice of measures is to check if the selected model actually models something that makes sense. This means that the model should have a relevant meaning to the substantive science.

4.2 General questions

The three questions 2, 6, and 10 asked about opinions on general topics or concepts, like the general consensus on minimizing model complexity, or the statement of George Box that all models are wrong or why we should pursue model selection in the first place if we do not believe in true models. The respondents were more

uniform in this area. They believed that a true model does not exist or is not attainable with finite data, at least when the meaning of 'wrong' is interpreted as that it is not the same as reality. The reasons are that truth is considered a static notion, while reality is more dynamic. The space of models is also considered very thin within the space of reality and models are considered only simplification of reality. An interesting remark was that a model that would be 100% consistent with reality would not be a model. There was one exception that felt that not all models are wrong since randomized experiments can go quite far in terms of giving causal interpretations, meaning that the 'truth' is almost attained. Furthermore, some respondents felt that models are formulated for a particular purpose and can therefore never be considered wrong; they can only be applied incorrectly. For instance, a hammer is not wrong, not even when you might need a screwdriver. Most of the respondents repeated what George Box added to his phrase, that some models are useful. Some claimed that it is better than nothing, since there is no alternative, but others were convinced that models are essential. A true model is not needed; the only thing that matters is a correct answer to the question. Even stronger, model selection is of importance especially when there is not a true model. Selection will help us find the best guess as good as possible. In case of a true model we would find it eventually. One respondent felt that model selection or model averaging makes sense only in textbook exercises, because when there is more at stake models are problematic. The question on model complexity was considered irrelevant by some respondents, since the measure of interest should completely determine the model selection, everything else would not lead to recognizable principles. Others felt that it would depend on the situation. If a simpler model would not include a relevant variable, then the simpler model would not be preferred. In this context, the measures of model comparison and complexity are only guides and not laws. On the other hand, simpler models have a higher chance of being applied and are easier to understand, which argues for simpler models. Some respondents viewed model complexity more practically. If a simpler model were to be more robust in prediction, simplicity would be preferred.

4.3 Purpose of model selection/averaging

The use of model averaging in relation to prediction and parameter estimation (question 5) was not clear to most respondents, since half of the respondents answered with 'don't know'. One respondent felt that models should not be averaged at all, but parameters should be averaged instead. Furthermore, comparing parameter estimates from different models should be avoided due to lack of independence. One respondent pointed towards the econometrics literature (without specific citations) that suggests that model averaging is good for predictions, but bad for parameter estimates. However, another respondent argues that for linear models model averaging would be beneficial both for predictions and parameter estimates, but this is less clear for non-linear models. The statement that 'AIC leads to the most predictive

model, whereas BIC leads to the true model' gave some controversy. One respondent argued that the question is completely irrelevant since most predictive is not well-defined and finding a model with the highest probability is a misguided goal. Others argued that the AIC and BIC do not take into account substantive relevance and could not agree to the statement. Of course, the respondents that do not believe in the existence of a true model could only agree to the first part, and some did on the basis of the definition of AIC.

4.4 Technical issues

An increased sample size is generally considered as a bonus, since it will generate more information, but there is some form of saturation since reality is not static and collecting data takes time. Thus while adding data, the true model may have moved on already. Reality may not be ergodic. Therefore, some respondents objected to this question since they do not accept the notion of a true model. The respondents seem to agree that whether including a variable known to be of relevance, although the data may not suggest this, depends on the situation. In some cases it is considered not necessary to include the variable since a simple misspecified model can do better in inference than a more relevant model, in particular if the model without the variable gives acceptable results. For random effects it can have other consequences though, since random effects are more sensitive to selection bias (e.g. in meta-analysis). On the other hand, some believe that adding just one such 'known to have effects' variable might not be too bad, but if many of these variables exist, the risk of over-fitting is present. A true Bayesian would of course include the variable, since all prior information should be included. Objective probabilities for model selection were considered unrealistic and cosmetic, and one respondent felt that this last part should be the exclusive domain of beauticians, hairdressers, etc. On the other hand, probabilities were considered epistemic, which does not mean that they are illusory. Model probabilities were considered useful guidelines, though perhaps not as laws. Furthermore, it was also believed that model selection can be based on expert judgment and generalizations of the modelled processes.

4.5 Open questions

The remarks from the respondents on the open question were quite diverse. Some liked the idea of the questionnaire and were interested in the results, while others felt that it has no use whatsoever. A specific question that came up was a question about the meaning of $P(\mathcal{M}_i \mid y)$. The answer should be that it defines the probability that model i is the best model to use for data based inference. It should not mean that it is the true or the actual model. Another question was whether model complexity is important in determining a good model. This did not receive a

conclusive answer. Somewhat related to this as well as to question 8, is a question about whether statistics should work in isolation from existing theories? In particular, how can the information criteria take advantage of existing theories? The answer was not provided, but the information criteria can easily be applied also to models that were selected on the basis of other theories.

5. Conclusions

In this article we have introduced the general frequentist and Bayesian ideas behind model uncertainty. The first is based on the Kullback–Leibler divergence, whereas the second is considering the posterior model probability. However, we have seen that in practice, the difference between frequentist and Bayesian model selection methods is probably smaller than antagonists on either side would like one to believe.

Some developments have not been discussed in this article. Boosting and bagging have entered the statistical vocabulary as ways to synthesize information using various models simultaneously. In the AIC and BIC literature these methods had their origin as Akaike weights or Bayesian posterior model probabilities, which can be used for subsequent inference. Again, similarities rather than differences are key, although the Bayesian paradigm is more naturally suited for doing post-hoc inference using these posterior probabilities.

Studying the answers of the workshop participants to the questionnaire demonstrated that the expected controversy in the field of model selection and model averaging is not resolved. For instance, there seems no uniformity in the choice on measures for model comparison, nor is there agreement on the use of the measures or the purpose of model averaging. True Bayesians answered the questions differently from those who do categorize themselves as frequentists. Some felt that model selection and averaging is textbook material and not for the real world. This controversy does not help in promoting the methods in real life settings. On the other hand, most respondents did feel that model selection and averaging is relevant for the quest of knowledge and believe that it contribute to science. As one of the respondents wrote ‘disagreement is a scientific principle of the highest order’.

Appendix A

This section contains the questionnaire that was sent out to all the participants and speakers of the workshop. In brackets it shows the number of people that selected that answer.

1. Why are you interested or think you have to be interested in model selection?
   - (a) It’s all pervasive (5)
   - (b) It relates to what I am working on (8)
   - (c) Other... (3)
   - (d) I don’t know (0)

2. The title of the workshop is "All models are wrong ...". What is your own idea about this statement?

   (a) Yes, all models are wrong (10)
   (b) No, not all models are wrong (1)
   (c) It is a bit more subtle (4)
   (d) I don't know (1)

3. How does sample size affects concepts such as "best model", "valid model", "useful model", and/or "reliable model". Is sample size fundamental in the discussion about whether models are always wrong?

   (a) With enough data the true model could be found (0)
   (b) Even with an infinite amount of data, the true model would prove evasive (11)
   (c) I don't know (2)

4. What kind of person are you? A model selection or model averaging person?

   (a) I am a model averaging person (2)
   (b) I am a model selection person (5)
   (c) I am neither (8)
   (d) I don't know (1)

5. Do you support the statement "I am happy with prediction through model averaging, but I have concerns about model averaging to estimate regression coefficients"?

   (a) Model averaging is good for prediction and good for estimating coefficients (2)
   (b) Model averaging is good for prediction and bad for estimating coefficients (5)
   (c) Model averaging is bad for prediction and good for estimating coefficients (0)
   (d) Model averaging is bad for prediction and bad for estimating coefficients (1)
   (e) I don't know (8)

6. If no true model exists, does it make sense to be doing model selection or model averaging at all?

   (a) Yes (12)
   (b) No (1)
   (c) I don't know (0)

7. Do you agree with the statement that AIC and cross-validation find the most predictive model, whereas BIC finds the true model (with high probability)?

   (a) I agree (4)
   (b) I don't agree (4)
   (c) It is a bit more subtle (5)
   (d) I don't know (2)

8. If you know a (random or fixed) effect of a particular variable exists, should you include it in your model?

   (a) Yes (3)
   (b) No (0)
   (c) That depends (11)
   (d) I don't know (1)

9. What is your favorite model geometry, i.e. the space (or distance) in (or with) which you like to compare models?

   (a) Kullback-Leibner divergence (4)
   (b) Bayes factor (2)
   (c) Coherence (no Dutch Books) (0)
   (d) Consistency (0)
   (e) Precautionary principle (0)
   (f) False positives (1)
   (g) False discovery rate (FDR) (2)
   (h) Other.. (3)

(i) It depends, because (2)
(j) I don't know (2)

10. In model selection it is often proposed or suggested that one should choose the models that has minimal complexity (e.g. less model parameters), if such models have similar fit or (AIC) performance as more complex models. Do you agree with this philosophy to minimize complexity when possible? If you agree how should we define and compare model complexity between non-hierarchical models which has similar fit performances?

(a) Yes (4)
(b) No (2)
(c) That depends (8)
(d) I don't know (1)

11. In one interpretation, model selection tools are aimed at determining a model whose predictions are aligned with true underlying probabilities. Do objective probabilities have to exist for model selection to make sense?

(a) Yes, statistical model selection is aimed at uncovering such chances (3)
(b) No, chances are illusory. We merely determine uncertainties pertaining to our beliefs (5)
(c) Well statistics is doing neither of these two things (6)
(d) I don't know

12. Feel free to write and answer your own questions

Note

1. This quotation, which translates to 'Plurality must never be posited without necessity', is most cited as the reason for the association of the ideal of simplicity with the historical Ockham.

References

Carnap, R. (1952), *The continuum of inductive methods*, Chicago, University of Chicago Press.
Chakrabarti, A. and J. K. Ghosh (2011), Aic, bic, and recent advances in model selection, in: P. S. Bandyopadhyay and M. R. Forster (eds), *Handbook of the philosophy of science, Vol. 7: philosophy of statistics*, Elsevier, London, pp. 583–605.
Claeskens, G. M. R. and N. L. Hjort (2008), *Model selection and model averaging*, Cambridge, UK . Cambridge University Press.
Dowe, D. L., S. Gardner and G. Oppy (2007), Bayes not bust! Why simplicity is no problem for bayesians, *The British Journal for the Philosophy of Science* **58**, 709–754.
Earman, J. (1992), *Bayes or bust?: A critical examination of bayesian confirmation theory*, Boston, MIT Press.
Eells, E. and B. Fitelson (2000), Measuring confirmation and evidence, *The Journal of Philosophy* **97**, 663–672.
Efron, B. (2004), The estimation of prediction error, *Journal of the American Statistical Association* **99**, 619–632.
Fitelson, B. (2006), The paradox of confirmation, *Philosophy Compass* **1**, 95–113.
Forster M. and E. Sober (1994), How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions, *British Journal of the Philosophy of Science* **45**, 1–35.
Forster M. R. (2000), Key concepts in model selection: performance and generlizability, *Journal of Mathematical psychology* **44**, 205–231.
Freedman, D. A. (1983), A note on screening regression equations. *American Statistician* **37**, 152–155.
Friedman, J., T. Hastie and R. Tibshirani (2001), *The elements of statistical learning*, **Vol. 1**, Springer Series in Statistics, New York.

Frigg, R. and J. Reiss (2009), The philosophy of simulation: hot new issues or same old stew? *Synthese* **169**, 593–613.
Gelman, A. (2011), Induction and deduction in bayesian data analysis, *Rationality, Markets and Morals* **2**, 67–78.
Green, P. J. (1995), Reversible jump Markov chain Monte Carlo computation and bayesian model determination, *Biometrika* **82**, 711–732.
Grünwald, P. D., I. J. Myung, and M. A. Pitt (2005), *Advances in minimum description length: Theory and applications*, The MIT Press.
Haenni, R., J. W. Romeijn, G. Wheeler, J. Williamson (2010), *Probabilistic logic and probabilistic networks, Vol. 350*. Dordrecht, Springer Verlag.
Hájek, A. (2007), The reference class problem is your problem too, *Synthese* **156**, 563–585.
Henderson, L., N. D. Goodman, J. B. Tenenbaum and J. F. Woodward (2010), The structure and dynamics of scientific theories: a hierarchical bayesian perspective, *Philosophy of Science* **77**, 172–200.
Hitchcock, C. and E. Sober (2004), Prediction versus accommodation and the risk of overfitting, *The British Journal for the Philosophy of Science* **55**, 1–34.
Kelly, K. T. (1996), *The logic of reliable inquiry*, New York, Oxford University Press, USA.
Kelly, K. T. (2007), A new solution to the puzzle of simplicity, *Philosophy of Science* **74**, 561–573.
Mallows, C. L. (1973), Some comments on Cp, *Technometrics* **15**, 661–675.
Morgan, M. S. (2005), Experiments versus models: new phenomena inference and surprise, *Journal of Economic Methodology* **12**, 317–329.
Rissanen, J. (1983), A universal prior for integers and estimation by minimum description length, *The Annals of Statistics* **11**, 416–431.
Romeijn, J. W.(2011), Statistics as inductive inference, in: P. S. Bandyopadhyay and M. R. Forster (eds), *Handbook of the philosophy of science, Vol. 7: philosophy of statistics*, Elsevier, London, pp. 751–775.
Romeijn, J. W. (2012), One size does not fit all: derivation of a prior-adapted bic, in: D. Dieks W. Gonzales, M. Stöltzner, M. Weber (eds), *Probabilities, laws, and structures*, Springer, pp. 87–105.
Romeijn, J. W. and R. van de Schoot (2008), A philosophical analysis of bayesian model selection for inequality constrained models, in: K. H. Hoijtink and P. Boelen, (eds), *Null, alternative and informative hypotheses*, Springer, New York, pp. 329–357.
Savage, L. J. (1951), The theory of statistical decision, *Journal of the American Statistical Association* **46**, 55–67.
Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics* **6**, 461–464.
Seidenfeld, T. (1979), *Philosophical problems of statistical inference: learning from R.A. Fisher*, Dordrecht, Springer.
Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B (Methodological)* **58**, 267–288.
Woods, J., Gabbay D. and S. Hartmann (2011), *Handbook of the history of inductive logic*, Dordrecht, North-Holland.

Received: 31 May 2011. Revised: 31 December 2011.